Research methods

È a process to generate knowledge

Knowledge: - the way through which something is presented to us (Kant)
- information structuring reality

the science of knowledge: epistemology

subfield: methodology
(= the science of “finding out”)

Linguistics as an empirical science:

- uses the empirical method: observation

Babbie (1998, pp. 9): “observation is the keystone of inquiry.”

Rules and Principles for:

- collecting
- evaluating
- interpreting data

= the standards of scientific research

Research methods are the strategies to achieve these standards.

Natural sciences vs. social sciences

- language: an artificial construct?

What should be emphasized - abstraction or experience?

Research incorporates both: deduction and induction.

È All sciences share the criteria of research.

The criteria of scientific research:

1.) objectivity
   • intersubjective verifiability (cf. Popper)
   • elimination of bias
   • documentation of data collection

2.) reliability
   • precision of collection
   • identical conditions yield identical results
   • data can be considered consistent and stable

3.) validity
   • strategy measures what it is supposed to measure
The nature of data
• measurable is the co-occurrence of phenomena
types of data relationships:
1. Coincidence
2. Correlation
3. Cause-effect relationship

1. Coincidence
• temporal relationship
• no inherent link
Examples: - "Titius-Bodesche Reihe"
- solar activity and stock prices

2. Correlation
• relationship between data that is non-causal
• consistent, stable relationship of variables
✦ can be: - linear (direct or indirect proportionality)
  - non-linear (can have max/min)
Example: grades in dependence on time spent studying

3. Cause-effect relationship
• cause temporally precedes effect
signal speed: $c$
effect of distance ("Fernwirkung")
✦ a causal relationship means: the removal of one variable stops /changes the event (in frequency, intensity).

The fundamental assumption of science:
• the physical universe is deterministic
• 3 cause-and-effect relationships
• these relationships can be analyzed

Data collection
• concerned with complex (linguistic) behavior
• data collection has 2 ends:
  - to establish the parameters of a system in preliminary studies, to get the extreme values
  - to establish benchmark data (benchmarks indicate how typical the data are)
Data:
- the formalized representation of facts, concepts, instructions
- suitable for further processing such as communication, analysis, interpretation

Ways to obtain data:
1. Observation
2. Measurement
3. Statistics

1. Observation
   • recorded data about phenomena, behavior
A natural observation:
   - data is taken as it is
     - in linguistics: non-verbal behavior, interaction research, interviews
B "guided observation":
   - in an experimental setup
     - in linguistics: ERP, PET, eye-tracking

Scientific consensus turns observations into facts

2. Measurement
   • numeric relationship between data should correspond to empirical conditions
   Example: number of prepositions in "exotic" languages
   forms in linguistic research:
     - surveys
     - interviews
     - tests
     - experiments

Error in measurements
   Interpretation of relationships: underlies error
   1.) measurement disrupts behavior
   2.) technological errors
   3.) individual variability
   4.) Rosenthal effect, Hawthorne effect
Error in measurements

Probability of an error:
1. accuracy (a measure of how close to reality an estimate is)
2. precision (a measure of how consistent a finding/estimate is)
3. replication (error when repeating the experiment)

3. Statistics
3.1. Descriptive statistics
3.2. Inference statistics

Inductive vs. deductive methods

Aristotle:

logic

observation

conclusion

Induction:
- preferred way of research since Renaissance
- 3 basic stages:
  1. collect relevant data; crucial point: observation process; facts are accumulated over time
  2. summarize data; employ systems of organizing/classifying data
  3. generalize about the system; generate laws

Inductive procedure

Flaws of the inductive method
- selecting the “relevant” data for analyzing
- probabilistic hence subjective conclusions
- other researchers draw different conclusions
- predictions have specific level of certainty
- judges from a number of observations:
  - from one observation (circumstantial inference)
  - from a limited set of observations
  - tendency to overgeneralize
- may be based on axiomatic premises
Deductive method

- required is an established body of strong generalizations

steps:
1. What to investigate
2. Hypothesis
3. Experiment
4. Verify/falsify hypothesis

Experiment

any systematic process that allows the testing of an hypothesis, involving the collection and analysis of data.

setup of an experiment = the experimental design:
- design specifies the components
  - type of experiment
  - samples to be studied
  - variables to be controlled
  - analysis type after collecting the data

Theory construction

HYPOTHESIS: a supposition arrived at from observation or reflection. leads to refutable predictions
- a tentative statement about a particular state-of-affairs
hypoetical constructs: imaginary entities
postulates: statements defining the hypothetical constructs and their relationships
coordinate definitions / operational definitions: translate behavior of hypothetical constructs into observable effects

Theory construction

THEORY: an abstract explanation for various laws/facts in an area of study
-a systematic structure of broad scope, encompasses a family of empirical laws regarding regularities in objects and events

- is internally consistent
- explains all/almost all phenomena in the field
- extensible and predictive; can lead to hypotheses
- has a certain predictive and explanatory power
- predictions made using the theory should be consistently accurate
Theory construction

Popper:

A theory can never be proven fully, one conflicting observation proves it wrong.

The deterministic view (Descartes, Kepler, Leipniz, Newton) established a new paradigm.

Kuhn: paradigm: the pattern underlying the construction of theories, explanations

Paradigms

replacement of paradigms: paradigm shift

natural sciences: classical vs. modern physics

linguistics behaviorism vs. Chomsky

Inference:

The scientific method is distinguished by its use of logic

inference = drawing conclusions from premises

fundamental logic argument: syllogism

Aristotle:
All human are mortal. Socrates is a human. Therefore, Socrates is mortal.

thus:
1.) if a is true, b will be true
a is not true. Therefore, b is not true. falsification

2.) if a is true, b will be true
a is true. Therefore, b is true. verification

Popper: agrees with 1.); denies 2.)
Inferences in social and natural sciences

- crucial influence: the **degree of predictability of humans**
- contrast of social sciences vs. natural sciences: corresponds with: free will vs. Determinism
- logic dilemma: we can only predict our decisions when we have met them

Descartes: dualism; the observer is separated from what he/she observes.

Popper: modern form of dualism

Popper: Wenn die physikalischen Gesetze dieser Welt autonom sind, sind wir nicht frei, wenn wir frei sind, sind die physikalischen Gesetze nicht autonom. (Logik der Forschung)

LAW: a statement describing a stable dependency between an independent and a dependent variable.

natural laws:

- descriptive vs. prescriptive
  - "soft" sciences vs. "hard" sciences

Criteria of social sciences:

- social regularities represent probabilistic patterns (regardless of counterexamples)
- social research deals with aggregated, not individual behavior
- social attributes: characteristics/qualities to describe persons (female, conservative, intelligent)
- social variables: logical groupings of attributes (gender, occupation, ...)

Scientific views in natural and social sciences

1. Empirism
   - knowledge as a collection of facts
   - universals are not obtainable
   - theories are summaries of observations
2. Operationalism
   - science is a system of rules
   - theories are tools for manipulation
Scientific views in natural and social sciences

3. Instrumentalism
- not the meaning of words is important but the way we use them
- theories are instruments of experience
- there is no "inner truth"

5. Realism
- laws have a relationship to reality that is relevant
- tool: observation

Linguistic methods: introduction

forms of linguistic description:
1. introspection
2. field research
output of 1.+2. generates models

MODEL: a representation of a process or system that shows the most important variables of the system in such a way that the analysis if the model leads to insights into the system.

Linguistic methods: introduction

models in linguistics
- core grammar of a language (e.g. Inuit grammars)
- category research (e.g. demonstratives in "exotic" languages)
- distribution of sociolinguistic variables (e.g. AusE - gender, origin (convict/non-convict), location (urban – rural - outback))
- computational models (PDP, connectionism)

Stances of linguistic research

1.) sociolinguistics
2.) descriptive linguistics
3.) generative linguistics

Language is an abstract object not amenable to direct observation" (Milroy, 1997)

Labov: - program of research for linguists begins with search for invariance best methods are developed by considering assumptions shared by linguists, not divided

landmark study:
The Labovian approach

- crucial concepts of linguistic research activity

1. subject of investigation (language, parts of language)
2. object of investigation (texts, recordings...)
3. product of investigation (model of language, grammar)

The Labovian approach

methods:

A.) the introspective method
B.) the analytic method
C.) the experimental method

A.) the introspective method

- based on self-observation, accessed by means of own competence
- no body of data
- disadvantage: variety or language not known cannot be studied
- intuition as a research aid

B.) the analytic method

- involves knowledge of target language
- generalizations based on a body of independent data
- example: discourse analysis

C.) the experimental method

- investigator "controls" data
- field research combines B.) and C.)
- entails manipulation of the informant's responses

Problems:
1.) can the structure of an unknown language be recovered from data analysis? (cf. Whorf, Malotki, Champollion)
2.) introspective methods may overlap (ex. Quirk/Greenbaum/Leech/Svartvik - combined analytic and introspective methods)
3.) most linguists would agree that data cannot be derived solely from a corpus
The generativist stance
- no corpus of language data, however large, can usefully serve as a basis for linguistic generalizations
- any corpus is partial and accidental
- replaces inductive with deductive research principles
- paradox in L1 child language acquisition
- led to establishment of psycholinguistics/cognitive linguistics

The generativist stance: paradox of psycholinguistics
L1 acquisition enables children to produce virtually infinite amounts of linguistic data.
Input includes:
- distorted input (also: deviant input; Chomsky)
- can be: mispronunciations, slips of the tongue
- omitted rules
- inference of rules out of defective material
- negative evidence = pointing at errors

typical errors in L1: *go-ed
atypical errors: *I no like syntax.

The descriptivist stance
- relies on phonological and morphological analyses
- basis of description: syntactic patterns of an unknown language
- assigns word classes to language material based on distribution, syntactic frames
- builds ranges of candidates for membership in linguistic categories

classic examples: Native American languages

problems: - reliability of speaker judgment
- investigator's bias

The sociolinguistic stance
- inductive research principles
- studies sociological aspects of language
- language considered in relationship with social roles
- goal: isolating the linguistic features used in particular, socially relevant situations
- distinguishing method:
  - sampling of speakers and language

Sociolinguistic methods: introduction
SAMPLING: process/method of isolating a representative group of individuals or cases from a particular population for purposes of drawing inferences from the analysis to the population as a whole.

goal of sampling: representativeness

On sampling
strategy:
1.) set the sample frame (list of population, phonebook etc.)
2.) define sample universe (i.e. set boundaries of groups of interest)
3.) assessment of relevant dimensions (includes social variables)
4.) fix sample size
On sampling

Sampling methods:
1. random sampling
2. stratified sampling
3. systematic sampling
4. quota sampling
5. judgment sampling

On sampling

Sampling methods:
1. random sampling
   • every member in the chosen sampling universe has an equal chance of being included
   • advantage: method is bias-free
   • disadvantage: sample may not be bias-free

On sampling

Sampling methods:
2. stratified sampling
   • relies on previous studies
   e.g. sampling universe includes: 50% speakers with English as a first language; 30% speakers with French and 20% bilinguals
   • random sampling from categories according to the ratios

On sampling

Sampling methods:
3. systematic sampling
   • every nth number after a random start is selected
   • e.g. sample 100 from a population of 1000: selects every 10th person
4. quota sampling
   • sampling plan includes certain pre-grouped numbers
   • e.g. 50% men, 50% women or: 50% over 18 years, 50% under 18 etc.

On sampling

Sampling methods:
5. judgment sampling
   • selects subjects arbitrarily
   • relevant: previous identification of types of speakers (e.g. includes only Cockney speakers)
   • characteristics of linguistic occurrence:
     - individuals
     - social groups
     - socially relevant situations

The cognitive change

Cognitive linguistics:
based on experience, perception, conceptualization
1. logical view (focus on syntax)
2. experiential view (focus on semantics)

Experiential view:
- “prominence” view on language
  - transfer of experience
  - conceptual tool: metaphor

11
The experiential view
prominence view: what is salient, what takes the attention.
experience: shared through same biological makeup
• physical experience shapes cognition
example: interpreting visual input
experience of moving objects does not relate to causes of motion
• interpreting visual input is an automated cognitive process

Meanings in grammar
are rooted in conceptual system of categories
categorization: combining a number of experiences into one conceptual directory
† creates conceptional categories
• are culture-related
• make sense only in distinction (cf. day-night)
Grammar reflects conceptualization.

Conceptualization in grammar
cf. gold nugget vs. gold dust – continuum of sizes
categorical distinction:
grammar codes gold nugget as a count noun, bounded
grammar codes gold dust as a mass noun, unbounded
Grammar does not reflect physical reality but experienced reality.

Conceptualization in grammar
Conceptual categories can be grouped vertically:
1. taxonomy
2. partonomy
• horizontally:
degree of typicality

Vertical grouping of categories
taxonomy partonomy

```
vehicle
car  bike
truck  van  convertible

body  wheels  lights
car
```

Horizontal grouping of categories
lexical and grammatical categories have:
1. prototypical members
2. peripheral members
Example: typicality of transitiveness
a.) Jennifer bought cherry yogurt.
b.) Jennifer loves cherry yogurt.
c.) Jennifer had cherry yogurt.

degree of transitiveness
On metaphors

- Cognitive abilities create new concepts out of existing concepts: mental transfer of concepts.
- Any physical, abstract, mental, social phenomenon is seen in the framework of a domain example: printed pages, reading, library (domain BOOK)

John is married to a library.

Incompatible domains (2-sided metaphor)

Metaphor: the mapping of a structure of one conceptual domain onto the structure of another conceptual domain.

On metonymies

2. The company wants to hire new brains.

Domain: human body

Metonymy: the substitution of one conceptual entity by another conceptual entity within the same domain.

On corpus methods

Corpus: body or collection of written or spoken material upon which linguistic analysis is based

- Used as a sample of language
- Provides view beyond individual experience
- Rules out individual salience
- Computer processable

Output:
- Concordances (i.e. lists of occurrences)
- KWIC (key word in context)
- Relative frequencies

Corpus search strategies

- Relative frequency of a word form: standard deviation from mean frequency of word forms
- Collocation: the appearance of one particular word form in certain distance of another particular word forms different meanings can have different collocates
  - Colligation: the appearance of one particular word form in a particular grammatical structure
  - Connotation: the semantic environment, can have positive or negative value

Corpus research

How frequent is a particular morphological form/grammatical structure?
Which particular structures have particular meanings?
Which particular structures have particular locations in texts?

- Corpus tasks have degrees of complexity

Relevance of tagging
**Types of corpora**

can be distinguished after:

1. **medium**
2. **purpose**
3. **research interest**

1. medium:
   - written (BNC - Aston & Burnard, Oxford)
   - spoken (London-Lund - Leech & Svartvik)
   - mixed (Birmingham corpus - Collins)

2. purpose:
   - child language acquisition/ learner varieties
     International Corpus of Learner English (ICLE) – Louvain
   - genre/topic specific corpora
     - sample only parts for special purposes
     - HCRC map task corpus
     - telephone communication corpus
   - multilingual corpora
     English-German translation corpus (Chemnitz)

3. research interest:
   - geographical varieties
     Brown (Kucera & Francis) – written AmE
     LOB (Johansson) – written BrE
     Macquarie corpus (Collins & Peters) – AusE
     ICE (Greenbaum) – international varieties
   - historical varieties
     Helsinki diachronic corpus (Kyto) – historical, dialectal English

**Corpus research**

empirical; questions introspection

different views:
weak corpus linguist's view:
  introspection and empirical research complement each other

strong corpus linguist's view:
  corpus data overrules introspection

- **research lines:** collocational, contrastive

**Collocations**

colloctional patterns are dynamic, change over time
Firth: ass – you silly ass
hoot – owl, car

implications of collocations:
- grammatical (colligation)
  Bolinger: look at vs. regard cf. She regarded him conspicuously. (with adverbs of manner)
- semantic: wreak = cause but negative
  (wreak havoc, wreak damage)
Contrastive studies

contrast between geographical varieties:
differences in use
BrE: greater frequency of modals, conditionals
AmE: more decisiveness

contrast between languages:
• use of parallel corpora
• aligned translations
can be: sentence-by-sentence
phrase-by-phrase
word-by word
Language contrast

OMC (Oslo Multilingual Corpus, Fabricius-Hansen)

- "sentence splitting": complex German sentences are split in Norwegian translations

Norwegian: 25% more singular sentences

Target text: more incremental

more coherence, anaphors

Wenn die unter jener anderen Landungsbrücke beheimateten Fische mein Kommen bemerkten, kam erschreckend aus dem Dunkel unter dem Steg hervor ein mehrere Meter breites und fast ebenso hohes und viele Male längeres, auf dem besonnnten Grund einen tiefschwarzen Schatten werfendes UNGEHEUER auf mich zugeschossen, das sich erst beim nahen Herankommen in eine Unzahl freundlicher Purpurmuguere auflöste.

Da min ankomst ble observert av fiskene som høyret hjemme under den andre landgangsbroen, skjøt det DET sjkrekkinnjagende UHYRE frem av mål under bryggen, og med kurs direkte mot meg! DETTE UHYRET var flere meter bredt, nesten like høyt og mange ganger sju langt, og kastet en dypsvart skygge på den solbelyste sjåbunnen. Først da DETTE SKREMSLET kom ganske nær meg, løste DET seg opp i et utall av vennlighetspurpurfisk.

Approaches in L2 acquisition research

1.) heuristic or hypothesis-generating
2.) hypothesis-testing

1.) heuristic: strategy of problem solving, aims at reducing problem-solving effort, usually involves incremental exploration of unknown terrain according to specific criteria.

Example: why are some language learners more successful than others?

- records observable factors; looks at behavior,
- tries to categorize it into patterns
- separation of verbal vs. nonverbal behavior

L2 acquisition research in the CING

hypothesis-generating
- data-driven
- no preconceptions
- inductive
- results: descriptions, hypotheses

hypothesis-testing
- hypothesis-driven
- use of predictions
- deductive
- results: theories
logfile         statistics

User ID: 0534
shared hReception-01.txt
Dauer load-exit: 16.163 s
focus (ms): 10465 0
blur  (ms): 0 0
focus gesamt: 10.465 s
blur  gesamt: 0 s
induktiv nCondIntroI.txt
load: Sat Oct  7 14:01:16 2000
exit: Sat Oct  7 14:01:22 2000
Dauer load-exit: 6.52 s
focus (ms): 0
blur  (ms): 0
focus gesamt: 0 s
blur  gesamt: 0 s
induktiv cond-tasks
Task 0 richtig
Task 1 falsch
Task 2 falsch

On quantitative research strategies
Types of quantitative research
1. quantitative linguistics a.k.a. statistical linguistics
   • sub-area of mathematical linguistics
   • controlled in investigation of ling. regularities
   • interested in frequency counts
   • for dictionaries, corpora
2. quantitative methods in socio- and psycho-linguistics

Language as a dynamic, probabilistic system
has certain integrity/stability
is combined from autonomy of their components
parameters of such systems:
   1. incidental/occasional
   2. regular
1+2 are the basis for modeling these systems by means of statistical distribution
distribution: most important characteristic of probability systems (example: weather)
distribution is an ordered aggregate of quantitatively expressed values

Quantitative linguistics
no specific body of concrete methods
logically grounded approach borrowed from other sciences
choice of adequate method determines outcome
used are: methods of quantitative mathematics
   probability theory
   information theory
aim: math. modeling of linguistic material
levels: linguistic engineering
   stylometrics (statistical stylistics)
   text typology research
   dialectrometry and language synergetics

Language as a dynamic, probabilistic system
logical basis: property and relation

\[ X \dashv Y \]

Text: X is connected to value Y on the basis of size (P)
ex. Hamlet \dashv 42.000
object # of NPs value
linguistic distribution: model description of language objects as probabilistic systems
Applications of quantitative linguistics

- grammatical category disambiguation
  - algorithms based on probability of co-occurrence of certain tags, reaches 96% accuracy
- “Sentence splitting”: complex German sentences are split in Norwegian translations
  - Norwegian: 25% more singular sentences
  - target text: more coherence, anaphors

Quantitative methods in socio/psycholinguistics

- separated through comparison to qualitative methods

Quantitative
- deductive - from theory to research
- tight research design
- control for (all) the variables
- focus on facts, relationships between variables
- modular approach
- data is presented using statistical results represented by numbers

Qualitative
- inductive - from research to theory
- modifiable research design
- no controls
- focus on context-specific relationships
- holistic approach
- data is presented using narrative description

Quantitative research design

should:
1. provide answers to specific research questions – results must have internal validity
2. control variance

Variance: the way and amount in which data or variables vary.

2 types:

a.) variance related to independent variables (= induced variance)

b.) unrelated variance (therefore requires control)

Relevance of control

The more components vary, the more difficult is control
aim is to reduce influencing variables by identification of the source of variance

Example: hypothesis: triggering L2 enhances language performance
- may differ individually (subjects may have been triggered before)
- cooperation problems
- problem of defining/measuring the dependent variable (language performance)
- “organismic” variables (indiv. health conditions …)
Relevance of control

If substantial variance occurs after intervention (i.e., experimental manipulation): safe to say that it is attributable to it (rather than to other variables)

† goal in controlling variance: separate variance attributable to indep. variables from that attributable to measurement error or extraneous variables

ways to control extraneous variables:
1. sample the subjects
2. introduce additional independent variables
3. keep conditions constant
4. make statistical adjustments

Methods of control

2. introduce additional independent variables
   takes "expected" extraneous variables into consideration
   Example: high-grade subjects vs. low-grade subjects
   † intervention may have different effects on them
   a.) subjects are randomized separately to gain even distribution or
   b.) make 2 groups to study the effect

3. keep conditions constant
   • coherent pre-test and after-test design
   • constant times etc.

4. statistical adjustment
   • change of data resolution - finer or rougher grid

Characteristics of quantitative research design

1. freedom from bias
   • the only systematic variation should come from the independent variable
   hence: goal is to eliminate bias and potential sources

2. control of extraneous variables
   • design must include procedures to minimize their effects

3. evaluation of quantitative research
   • through statistics, design should produce data for all stated hypotheses
   • precision of data collection should enable hypotheses to be tested with confidence

Qualitative research strategies

approaches:
1. language oriented: meaning of words, social interaction, communicative function of lexical material

2. descriptive/interpretative: description of linguistic phenomena and of their speakers’ environment

3. theory-building: identifies connections between phenomena
   quantitative data: concerned with numbers,
   qualitative data concerned with meanings
   practice: both approaches complement each other
   Ex. statistical analysis of qualitative interview data – HRT in AusE

Nature of qualitative data

• unstructured, multi-medial
• subject-defined, not pre-structured
• types of data depend on method of collection

What is the main difference between the students of the 1980s and the 1990s?
† 30 years

What result would you get if you laid a class of 30 students, average height 5’5”, end to end?
† They would all fall asleep.

numbers and meanings: related at all levels of analysis

Nature of qualitative data

• concepts base on limited numbers of observations
• numbers base on meaningful conceptualizations (example: pattern recognition in outback talk – numbers of occurrences of diminutives)
• meaningful conceptualizations are informed by numbers
   (example: hypothesis about functions of diminutives: confirmed or refuted)

qualitative analysis: resolving data into constituent components

(Examples: Dey 1995, 15)
Qualitative data analysis

Steps:
1. comprehensive description of phenomena
   - “thick” vs. “thin” descriptions (Geerz et al)
   - context
   - intentions
   - subsequent evolution

2. recognition of context
   - social contexts: institutions, societies, cultures
   - spatial contexts: urban settings, rural settings
   - can vary over time
   - can provide key for meanings (cf. anaphora, ellipses)
   - meanings: context-dependent

3. intention as a factor
   - focus of qualitative analysis: describing reality as perceived by different observers
   - meanings: negotiable between different observers
   - interference of meanings with individual factors
   - relevance of motivations

3. classification of data
   - establishing a conceptual framework for data
   - basis for interpretation
   - breaking up and rearranging of data
   - classification should always be guided by research objectives

4. identification of connections
   - finding associations between variables
   - starting point: logical connections (example: gender distribution in AusE)
   - pattern recognition within data

Survey research: oldest technique in social data collection; cf. Old Testament

After a plague the Lord said to Moses and to Eleazar the son of Aaron, the priest, “Take a census of all the congregation of the people of Israel, from twenty old and upward...” (Numbers 26:1-2)

Ancient Egypt: census

today: most frequently used method of observation in social research
surveys used for: descriptive, explanatory, exploratory purposes

Interview surveys

- alternative way to collect data
- questions asked orally, answers recorded
- face-2-face or telephone
- small-scale interviews: one interviewer

advantages:
- higher response rates than mailed questionnaires
- less “no answer given” fields
- guard against confusing questionnaire items
- interviewer records and observes
Interview surveys
- assumed: question items mean the same for all respondents
- response must mean the same from different respondents
- interviewer’s presence should not affect the perception of a question or render specific responses
  interviewer: should be a neutral transmitter medium

general rules for interviewing
- appearance and demeanor: should be adapted to subjects/respondents
- cultural awareness (ex. Likert scale in Appalachian: very means fairly, rather, poorly)

Theoretical surveys
- theory-construction upon previously collected linguistic materials
- used for studies affecting entire languages rather than varieties
- linguistic universals surveys
- includes data from different fields of study, e.g. corpus materials

cf. Diessel on demonstratives
one-term systems (German: Das Haus da. French: Cette maison-là. Swedish: Det här huset.)
two-term systems (English: this/that)
three-term systems (Spanish: este, ese, aquel, four-term systems (Hausa: na, nà, cân, can)
five-term systems (Malagasy: ily, io, ira, ira)

Theoretical surveys

Questionnaires - concepts and design
operationalization: development of specific research procedures - in social sciences: asking questions
questionnaires: used in surveys, experiments, field research
interview vs. questionnaire: depth vs. breadth
questionnaire features:
- self-administered
- has questions and statements
- statements: for determining extent of attitude or perspective of the respondents

Questionnaire design
1. format
   - “spread out and uncluttered” (Babbie)
   - avoid abbreviations, narrow spacing
   - subsections should be numbered

2. adaptation to specific users
   - contingency questions = subsequent questions that apply only to some respondents
   - ex. If 4. applies, please specify.

3. answering efficiency
   - issue of length (Barrat/Cole: 10 minutes)
   - use of matrix questions
   - Likert-scale questions

Relevance of questions
open-ended questions vs. closed-ended questions
open-ended: must be coded for analysis
danger of irrelevance
closed-ended: can be transferred directly
disadvantage: omissions can be in the checklist

answer categories: mutually exclusive
  exhaustive (i.e. all possible answers covered)
to overcome interference:
  -category answered should fit best
  -unambiguous, simple
  problematic: coupling of questions:
  Do you agree that dialect has a lower status but a higher function among the peer group?
Questionnaire design

4. order of questions
• difficult / most “interesting” questions first (Barrat/Cole: easiest questions first p.103)
• previous questions can affect later answers
  ex. closed questions provide set of answers, more likely to appear again in subsequent open-ended questions
• questions may be randomized, but:
  internal “logic” preferable

5. instructive
• questionnaire should begin with basic instructions
• specify extent of open-ended questions
• instructions how to return the questionnaire

Likert scaling in questionnaires
• items may have intensity structure or continuum
  Likert scaling: systematic construction of indexes from questionnaires
  ex. questionnaire presents a statement.
  Likert scale = strongly agree – agree – disagree
  – strongly disagree – undecided
  + assigns numbers to scores

Bell’s checklist
a.) How long did it take to complete the questionnaire?
b.) Were the instructions clear?
c.) Were any of the questions unclear or ambiguous?
d.) If so, which, and why?
e.) Did you object to answering any of the questions?
f.) In your opinion, has any major topic been omitted?
g.) Was the layout of the questionnaire clear/attractive?
h.) Any other comments?

Statistical methods in linguistics

descriptive: orders, summarizes data
inferential: assists in drawing conclusions from observations

† descriptive statistics is a method for presenting quantitative data in a manageable form

• exhaustive investigation:
  theoretically impossible: to observe any given string in a natural language
  theoretically possible but not feasible: observe any speaker of a language
  -necessity: data reduction
  -recognition of trends in data - frequency distribution
  -determination of errors - probability theory

Frequency analysis

1.) data is ordered from highest to lowest value
  e.g. coding of a Likert scale on English proficiency
  rudimentary - poor - average - good – excellent
  or: points in placement tests 0 to 60

2.) make arrays/intervals:
  1  0 to 10 rudimentary
  2 11 to 20 poor
  3 21 to 30 average
  4 31 to 40 good
  5 41 to 50 excellent
  † creates a frequency table

-frequencies order data, enable the calculation of statistical values

Frequency analysis

1. correlation
2. measure of association
3. central tendency
4. variability

1. correlation
• correlation coefficient r: strength/direction of a relationship between 2 variables
• to what an extent two features are related
• chaotic relationship: r = 0
  r = +1 for perfect proportionality
  r = -1 for perfect reversed proportionality
• type of r used depends on linear or a nonlinear relationship
2. measure of association
PRE: proportional reduction of error

<table>
<thead>
<tr>
<th>age group</th>
<th>0-19</th>
<th>20-85</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>deductive</td>
<td>50</td>
<td>500</td>
<td>600</td>
</tr>
<tr>
<td>inductive</td>
<td>300</td>
<td>100</td>
<td>400</td>
</tr>
<tr>
<td>total</td>
<td>350</td>
<td>650</td>
<td>1000</td>
</tr>
</tbody>
</table>

2.1. overall distribution is unbalanced
+ prediction would generate 400 error cases
- based on age scale generates 150 error cases

\[
\Lambda = \frac{250}{400} = 0.625
\]

theoretical \( \Lambda = 1 \): perfect statistical association

3. value of central tendency
= one value represents a group (e.g. for comparison)

3.1. modal value (M)
= value that occurs most frequently in the survey
disadvantage: not really central, there may be a few values smaller and many values higher

3.2. median (mdn)
- separates all values in half (50% of data is higher, 50% lower)
- advantage: not influenced by extremes

3.3. average or mean (m, \( \bar{x} \))
\[
\frac{x_1 + x_2 + \ldots + x_n}{n} = \bar{x}
\]
-disadvantage: extremes matter

4. value of variability

4.1. measures for variability:
variation range = difference of \( x_{\text{min}} - x_{\text{max}} \)
-disadvantage: simple, may depend on coincidences, extremes

4.2. standard deviation SD
= the scattering around \( x \)-bar
tells how variable the data are around \( x \)-bar
Gaussian distribution: on either side
68% of values lie within one SD
95% of values lie within two SDs
(large SD: less values are close to \( x \)-bar)